DOCUMENT RESUME

ED 346 106 TM 017 873

AUTHOR Guion, Robert M.
TITLE What I Wish I Knew about Assessment.
PUB DATE Jun 91
NOTE 24p.; Paper presented at the Annual Meeting of the International Personnel Management Association (Chicago, IL, June 24, 1991).
PUB TYPE Viewpoints (Opinion/Position Papers, Essays, etc.) (120) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Decision Making; *Employment Practices; *Evaluation Methods; Evaluators; *Knowledge Level; *Measurement Techniques; Personality Traits; *Personnel Evaluation; Personnel Management; Personnel Selection; Position Papers; Predictive Measurement; Program Evaluation; Standardized Tests

ABSTRACT

Reflecting on a career spent in assessment in personnel selection and managment, the author lists 12 things (items) that he wishes he knew about assessment. The first four items are underdeveloped ideas set aside because of the field's preoccupation with equal employment opportunity. A second set (items five through eight) comes from intellectual neighbors in other disciplines. Items 9 through 12 grew from pure frustration. The author wishes that he knew: (1) how to handle predictions of different criteria when predictions are inconsistent; (2) how to individualize assessments in making employment decisions; (3) what personality variations to assess for what kinds of performance problems, how they might be assessed, and what their assessment would add to the accuracy of predictors of performance or other job-related behaviors; (4) how to select teams or replacement members of existing teams; (5) what a "standardized" test really is; (6) whether excellence in measurement is really much better than excellence in non-metric assessment; (7) more about test-taking strategies; (8) more about program design and evaluation; (9) whether simulations and related content assessments really work and aid traditional aptitude assessments; (10) more about the physical demands of jobs; (11) why it is no longer assumed that decision makers have brains; and (12) how to handle those who do not recognize the need for improved assessment. Included are two tables, an eight-item list of references, and one flowchart. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
ROBERT M. GUION
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED345706

WHAT I WISH I KNEW ABOUT ASSESSMENT

Robert M. Guion
Bowling Green State University

A quarter of a century ago I wrote a book on personnel testing. Currently, I'm writing another on assessment for personnel decisions. Twice I've done Annual Review chapters with sections on assessment. For many years, I did individual assessments for promotion, transfer, and hiring decisions. I've developed and validated testing programs and exercises for assessment centers. I've had the expert witness experience. I've been invited to speak -- here and elsewhere -- on topics related to personnel assessment procedures, presumably because the people who invited me thought I might have something valuable to say. I've thought so, too.

Now I'm not so sure. In retirement, one wants to look back fondly at what one has accomplished. It's unsettling to realize that I -- forgive me, but I'm changing that to we -- that we haven't accomplished nearly as much in the last quarter of a century as all the activity would suggest. It becomes particularly unsettling when one is trying to write a book that, with a little bit of good sense and a whole lot of good luck, might have something to say about what we should all be doing in the next quarter century. It is here that I've run up against matters in which I have relatively little experience, in which most of my friends and colleagues have equally little, but which seem extremely important and overdue for serious consideration.

Presented to the IPMA Assessment Council at its annual conference, Chicago, June 24, 1991.

BEST COPY AVAILABLE

TM017873

The last quarter century can be identified as the EEO era in personnel selection. A little bit before the EEO era began, Marvin Dunnette outlined an idea for individualized assessment and decision making (Dunnette, 1963a). A little before that, he and I and others wrote articles calling for the prediction of multiple criteria (Dunnette, 1963b, Guion, 1961). Such ideas and others of the early 1960s were worth pursuing. We didn't pursue them because we got preoccupied with laws and regulations and court decisions that, in all too many respects, froze the field of personnel assessment where it had been.

I make no apology for that preoccupation. Our society, through its many prejudices, had placed too many obstacles to good jobs in the way of too many well qualified people. We -- our society -- needed to remove those obstacles, and we -- professional assessors -- needed to use our techniques, with the help of enabling law, to help remove them. Many private employers abandoned testing, but I believe our methods and rules of evidence had a positive and important influence in removing those obstacles. Nevertheless, I regret that the unfolding of events early in the EEO era fossilized some ideas and procedures that deserved further development. I wish I knew how that development would have turned out.

Not everyone in the assessment field was so directly preoccupied with EEO issues. The coming of the computer age led scholars to important advances in educational assessment methods and to new areas of research in cognition. Some of the things I wish I knew are relatively new developments in these other fields, and they have untried relevance

for us. We might now know much more about them and their value for employment procedures if we had not set them aside to play lawyer.

I've settled on a list of twelve things I wish I knew. Some of them call for research, but I've never accepted the notion that researchers and practitioners are different folks. The list is incomplete, but a list of twelve is enough to stretch your patience. They can be grouped under three headings. Those in the first set stem from the undeveloped ideas set aside because of the EEO preoccupation. The second set comes from the work of our intellectual neighbors. The third set -- a nontrivial set -- has grown out of plain frustration!

The Underdeveloped, Frozen Ideas

1. In 1961, I suggested a five-step validation sequence that called for the identification and valid measurement of multiple criteria, and the validation of predictors independently against each of them (Guion, 1961). Specifically, I argued that the relative importance of different criteria was not important at the time of validation; it becomes important at the time that administrative decisions have to be made. That is, the relative importance of predicted quality versus quantity may depend both on the relative validities of those predictions and changes that may occur in the productivity history of the organization. Marv Dunnette had very much the same idea: He said, "Let us cease searching for single or composite measures of job success and proceed to undertake research which accepts the world of success dimensionality as it really exists" (Dunnette, 1963b, p. 252). We may both have been guilty of excessive fractionation of the idea of job success, but there are different kinds of criteria. Performance quality is different from staying or quitting.

Consider this table:

Probability of:	Candidate		
	A	B	C--
Above Average Performance	.80	.20	.70
Quitting Within Year	.10	.40	.80
Being Fully Trustworthy	.90	.85	.35

Candidate A is an obviously good choice; every prediction is consistently good. This is a candidate who is likely to perform well, be trustworthy, and not quit. B is likely to be trustworthy, and may stick around, but is not likely to be an above average performer. C is likely to perform very well indeed, but there's some question about C's staying power or trustworthiness. What decision will you make about B or C?

Wish #1: I wish I knew how to handle predictions of different criteria when the predictions are inconsistent. It shouldn't be too hard to work on. Standard policy-capturing might be used to develop an equation, or perhaps a set of equations, to model the judgments of organizational decision makers; lens model research could be used to validate those judgments. But against what? The lens model uses a single criterion!

Neither Guion nor Dunnette nor anyone else developed the idea in the early years of Title VII. We faced a big enough problem trying to explain to lawyers, enforcement agencies, and confused employers how to validate with even one criterion. If we had been more successful with that, we might have worked on the case of inconsistent multiple criteria. But we didn't, and I wish I knew what we would have learned had we done so.

2. Dunnette also presented a model for individualized selection (Dunnette, 1963a). The model recognized two unfortunate facts, too often ignored. One is the assumption in traditional validation that everyone is just like everyone else on everything except the predictors and the criteria, and that therefore the same predictors and regression equations should be used for all. The other is that we have no research model for validating predictors with an n of 1. Dunnette's very intelligent point was that we should establish homogeneous groups of predictors, of people, of work situations, of job behaviors, and of organizationally-relevant consequences of those behaviors in those situations. That is, he proposed a massive tree of moderators such that predictors used for one group of people, to predict one kind of behaviors, might prove different from those used for other categories of people and behavior. Should we use the same kinds of assessments for applicants whose backgrounds suggest that they could learn to do the job and for those who have actual experience in doing it? I would argue for aptitude tests for the former and work samples for the latter.

With gross over-simplification, Figure 1 gives the flavor of the Dunnette model. Suppose we have an accepted but unarticulated policy against hiring people who ignore no smoking signs, are dirty and unkept, spit on the floor, or talk abusively to others in the waiting room. Our first basis for decision is such a policy. If the candidate doesn't fit the policy, some euphemism is offered by way of rejection. Otherwise, the candidate gets interviewed. In this simple chart, the interview seeks only two kinds of information: What is the candidate's general history of prior accomplishment, and does the candidate have actual experience in the job at hand. If the history includes many impressive

accomplishments in work, school, or community life but no experience in this specific job, the person might be hired and trained or considered for some other position; we would not let real gems get away. If such a candidate does have relevant experience, or if another person has a modest record of accomplishment plus real job experience, we might administer a work sample test. A high score is a sure further reason for hiring the person of many accomplishments, but we might insist on a really high score for the one with more modest prior achievements. Poor work sample performance would tilt the decision toward rejection. With someone with no record of achievement, even a lot of relevant job experience might not outweigh the sadness of the record; such a candidate might be rejected. People with little or no achievement history and no relevant job experience could be tested to assess both aptitude and work motivation. If their scores are good, we hire them; with poor scores, we reject them.

Nothing can be firmly asserted about this flow chart other than that it is one -- and that it does not represent a conventional multiple regression procedure based on all available cases. Some people don't even get the preliminary interview. Of the others, some are tested, some do work samples, and some get hired without any sort of further assessment. Whatever else may be said, whatever questions we would ask about this procedure and the jobs it would fit, it might represent employment office reality better than our usual statistical models do.

We were never very happy with treating everyone alike. After all, the whole concept of selection is based on the fact of individual differences. Dunnette's model might have been taken seriously had we not been sidetracked in treating demographic variables -- protected

classes -- as potential moderators. They weren't, so we said there are no moderators.

Wish #2: I wish I knew how to individualize assessments in making employment decisions, in a way something like Dunnette's model. I suspect that there are psychological, nondemographic moderators -- perhaps in attitudes, background experiences, test-taking strategies, or temperament -- that would make the model work if the search for them were given some serious thought. I suspect that there are good statistical models for evaluating the components of the process -- but I don't know them. Maybe some sort of path analysis or other structural model might work, but my impression is that everyone in such an analysis must follow a common path.

I think the model requires a lot of background research. The taxonomic requirements alone will take years to develop. When the various categories are satisfactorily in place, many bivariate studies will be needed, followed by meta-analyses, to form the required linkages.

3. According to an old axiom in personnel work, performance is a function of ability and motivation. Maybe motivation is provided by good management, but many of us have believed, and I still believe, that the interests and personality traits people bring with them in applying for a job can to some degree add something to the prediction of performance.

Dick Gottier and I reviewed research on personality test validity in 1965. We did not, as so many have assumed, conclude that personality tests are no good; indeed, we found evidence that "in some situations, for some purposes, some personality measures can offer helpful

predictions" (Guion & Gottier, 1965, p. 159, emphasis in original). What we did conclude was that (a) research designs were consistently inadequate, (b) raw empiricism resulted in better prediction than did existing theory, and (c) that there was no generalizable evidence of the value of personality tests in employee selection. Had we not turned our attention almost immediately to the dead end of test fairness, I suspect that we would have followed up on these conclusions with better designs, with theories explicitly tied to the work motivation aspects of personality or to specifically relevant criteria, and, in time, reached some generalizations about what personality variables were associated with definable job behaviors. As it is, however, Wish #3: I wish I knew what personality variables to assess for what kinds of performance problems, how they might best be assessed, and what their assessment would add to the accuracy of predictions of performance or other job-related behavior.

4. In the 1950s, the Air Force did a substantial number of studies of air crews and how they might best be formed. Group processes and structure were extensively studied in experimental social psychology and some private employment settings. Some of the relevant research centered on the initial formation of teams and on selecting new members for existing teams.

It is less than accurate to blame the decline of such research on the advent of the EEO era, but EEO concerns were surely relevant. Consider one nonmilitary example. In 1952, Van Zelst (1952) allowed some carpenters to form work teams according to their own choices; other work teams were formed by assignment. The groups formed by sociometric choices had lower labor and materials costs, less turnover, and greater

job satisfaction. Now consider these findings in the light of the post-1965 era, with a mixture of ethnic groups and some women in the pool of carpenters. How integrated would such self-formed groups be? We don't really know, and we don't know much about the trends that might have been observable in self-formed groups, but we can be forgiven for thinking that such groups would be homogeneous in more than ability.

Another practice common in the 1950s was individual assessment of candidates for managerial positions. A typical procedure included a visit to the employing organization to find out something about it and to reach some conclusion about the characteristics of people who would "fit in" or "be a team player." Then we talked about fitting into the existing network of interpersonal relationships; today we would talk about fitting into the organizational culture. But we have some queasiness about this notion of "fitting in."

Wish #4: I wish I knew how to select teams, or replacement members of existing teams. I wish I knew how to talk about hiring people who will "fit in" without worrying that the phrase itself may mask a desire to discriminate on grounds unrelated to performance.

The Ideas from the Neighbors

5. Consider the phrase, standardized testing. It has an egalitarian ring to it. Everyone is assessed in a standard way, with no hint of favoritism for anyone. We used to talk about two kinds of standardization, the standardization of score interpretations and the standardization of procedures. Everyone was given the same items, with the same time limits, the same size type, the same response options, and the same scoring key, and the score was interpreted in the same way -- either from a norm table or an expectancy chart.

With the advent of item response theory and computerized adaptive testing, things changed. Different people answer different items and even different numbers of items. A score is described not with a percentile rank or a probability of success but with an estimate of ability called theta. It brings me to wish #5: I wish I knew what a "standardized test" is. I'm prepared to say that the testing procedure, and certainly the score interpretation, is far more standardized with IRT than in more traditional testing because it is the logic of the IRT algorithm of measurement that is standard. I am not prepared, however, to say that measurement by any other standard algorithm (for example, Guttman scaling of difficulty levels) would have the same egalitarian flavor. So, at a very basic level, I have to admit that there is something about the idea of test standardization that has eluded me -- and continues to elude me.

6. I distinguish between the use of the word measurement and the word assessment. Assessment is the more global, more inclusive term. It includes measurement, but it also includes more qualitative evaluations of more sweeping characteristics of people. I refer to measurement when I use a procedure that assesses a rather specific trait, one that is in some sense unidimensional so that people can be ordered along a single scale. I refer to assessment when I use a procedure that assesses a kind of general fitness for some purpose, when fitness can be defined perhaps by a variety of different combinations of relevant traits.

Over the last two or three decades, psychometric research has enabled us to do far better measurement than we typically do. To take full advantage of the advances that have been made, we need to define a

trait to be measured very carefully, to identify the behaviors or signs of the trait as defined and the behaviors or signs of other traits that might get confused with it, set up generalizability studies to determine the operational limits of the intended score interpretations, calibrate items using a one-, two-, or three-parameter item response model, and select the items that will yield the most precise possible score for any given examinee.

Wish #6: I wish I knew whether excellence in measurement is really much better than excellence in non-metric assessment. In part I'm wondering whether very precise measurement of one or two well-defined traits is better than carefully developed methods of global assessment. Stated differently: I wish I knew whether maximizing measurement validity is a better strategy for personnel decisions than maximizing job relevance. If you are among those who still equate validity and job-relatedness, consider the use of an examination with a very high level of construct validity as a measure of inductive reasoning for a hamburger-flipping job in a fast food palace. Validity in this sense is no assurance of job relevance. My concern is that, in seeking only psychometric elegance, we may be failing to study and to understand real job requirements. Does a sophisticated method like IRT help us do a better job of selecting employees than a home-made set of interview questions that are essentially an oral test? It might -- if the trait measured is truly important for the work at hand, or if the interview is as haphazard as most. But is it necessarily better? I wish I knew for sure.

7. Let's back up to the EEO problems and moderators. It has been established that demographic variables are unlikely moderators of

employment test validities. Nevertheless, there is the disquieting fact that racial mean differences in test scores are greater than those in performance. Regression to the mean may account for this, or at least part of it, but I'm not wholly comfortable with that explanation. Moreover, I'm struck by what seems increasingly to be an approaching bimodality in minority performance on both sides of the prediction equation. That is, in the last quarter-century, many minority people have moved into middle class living with all it implies -- better housing, education, jobs, and general experiences -- and greater comfort with middle class American culture. Many others seem permanently enmeshed in poverty. Does a person's economic status tell us anything useful about the validity of that person's test score? That is, is the score a better description of ability for people with "advantages" and weaker for those called "disadvantaged"? Yes. Although various environmental scales failed to serve well as moderators, we know that test scores tend to be somewhat related to socio-economic status. But why?

Studies of differential item functioning have considered one possible aspect of the answer, but they are usually based on the demographic categories for purely statistical reasons. Maybe we can do better. Cross cultural research has grown in the last couple of decades. More serious attention has been given to the equivalence of measurement instruments in different languages and for people of different cultures. There is reason to question whether the meaning of taking a test is constant across cultures

Wish #7: <u>I wish I knew something about test-taking strategies</u>. I would like to know if people with a middle class background, people who

have gone through reasonably good schools and have taken it seriously, approach a test differently from those from a poverty culture whose educational background is weak or abbreviated. Are they more likely to guess and move on or to leave an item blank for later consideration? Are they more or less likely to change responses? Do they spend more time on individual items? Do they tend to rule out certain distractors before guessing among the others? Is it not at least conceivable that test scores on the average are better for people who follow certain strategies than for people who use other strategies? And is it not at least conceivable that, in certain unassimilated ethnic or cultural or economic groups, nonoptimal test strategies are characteristic? I've heard people talk about test-taking strategies for years, but I've seen little in the way of relevant data. One reason is that traditional studies of strategy have had to rely on retrospective comments. I believe that computer technology can help us do such studies far more effectively. In any case, we ought to be working toward some way to identify what it is within an ethnic or cultural group that results in a bigger mean difference in test scores than in criteria.

8. For jobs where lots of people are hired, traditional methods of validation work quite well. Increasingly, however, organizations are small and hiring is limited. A position, not a multi-position job, comes open and someone must be hired. The position may be unique in important ways. Someone else must be hired for an unrelated position. Do we have anything to offer employers in such settings? I think so. Some validity generalization information may be helpful. Structured interviews may help. What might help most, however, could be a systematic program in which standard procedures are used to identify

position responsibilities, traits to be assessed for designated responsibilities, and the feasible ways of assessing those traits. In such a system, one would evaluate, not inferences from test scores, but the program as a whole.

Comprehensive programs are evaluated by quasi-experimental research. Here are two very simple designs:

Design 1:

Time 1	Time 2
Program Used in Some Locations	Mean Performance Observed
Program Not Used in Control Locations	Mean Performance Observed

Design 2:

Time 1	Time 2	Time 3
Performance Measured in Group 1	Program Used	Performance Measured
Performance Measured in Group 2	Program NOT Used	Performance Measured

Design 1 has been used several times in validating integrity tests. In a chain store, for example, the test may given in some experimental group of locations and not in other, somewhat similar control locations. After time, the criterion performance (usually inventory shrinkage) is measured. If there is less shrinkage in the stores where the test was used, the selection system is considered valid.

This really is a rather weak design. Not much can be said about causes. Is the test valid, or is shrinkage reduced only because the test shows corporate concern about employee theft, or were the two sets of locations poorly matched? Design 2 is still weak, but it's a little bit better because it calls for measuring shrinkage before doing any testing in any of the locations. If the control locations show little or no change, and the experimental locations show substantial improvement, problems with Design 1 are reduced.

Wish #8: I wish I knew more about program design and evaluation. Cook and Campbell (1979) described many different quasi-experimental designs and some of them are stronger than others; most are stronger than these. Program evaluation methods have been around a long time; their popularity has both waxed and waned without my ever being involved in the debates. I don't know much about their virtues or weaknesses, and I'm not at all sure how to use them for evaluating a selection decision program. Indeed, I'm not sure what to include in an integrated assessment-and-decision program. I'm not alone. For most of us, program evaluation hasn't seemed especially relevant to personnel selection. We'd better take another look.

Frustrations

9. Many of us believe that actual performance on realistic job simulations offers a better assessment of candidate qualifications than performance on multiple-choice aptitude tests. Some jobs are unique; candidates cannot have had any experience on the job, so realistic simulations may not be feasible. In one situation I know about, a simulation exercise was developed to follow a rather elaborate, and I suspect good, abstraction of a training program. The work sample

exercised was developed to see how well people actually perform the newly learned tasks that are small but prerequisite parts of the real job. A great deal of research time and money has gone into this project, certainly far more than would have been required for the multiple-choice aptitude test.

In an unrelated activity, I've been serving on a panel to make some recommendations to a federal agency. The rest of the panel is quite convinced that the traditional multiple-choice test is a bad idea and is recommending "new" -- and untried -- approaches to assessment. The new ideas being suggested are indeed intriguing, and I hope the agency well.

But, wish #9: I wish I knew whether simulations and related content-oriented assessments work as well as we think they do, and if so, whether they work as well as the more traditional assessments of aptitudes or whether they add incrementally to aptitude assessment. We do need to work diligently to improve what we can do, to develop new constructs more relevant to work performance, and to develop alternative and preferably improved ways to measure them. But we must avoid the trap of assuming that new is necessarily improved, and I'm getting terribly frustrated by the increasingly common attitude that enthusiasm for the new -- whether the once new idea of assessment centers or the currently new notion of triarchic theories of intelligence -- are necessarily better without confirming comparative evidence.

10. It has been very frustrating to work with colleagues who consider cognitive abilities the only ones worth assessing. Sensory abilities, motor skills, and physical condition -- these are almost universally ignored or, if considered, considered and assessed

superficially. Except for the fitness center and the jogging path, muscles are simply no longer in style. The image of a job requiring the big he-man with bulging muscles is pretty much limited to work with a jack-hammer. For many years, industrial engineers and miscellaneous inventors worked to find ways to simplify jobs, making it possible for wimps to do the work that formerly required Charles Atlas types. The Civil Rights Act accelerated the trend by making it useful to find work aids to help women do jobs that previously required high levels of strength. The Act for Disabled Americans will call for further innovations in accomodating to a variety of physical problems.

Indeed, it is their variety that concerns me. The physical demands of jobs are not merely muscular. Susceptibility to allergic reactions or to physical stress or even physical susceptibility to effects of emotional stress are physical problems associated with work. Wish #10: I wish I knew more about these kinds of physical demands of jobs. They are important not merely because of the laws and the potential for litigation under various legal theories; they are important because some effects may cause long term problems for employees -- problems more long lasting than a sore muscle and requiring treatment more sophisticated and expensive than liniment.

11. A major frustration for me these days is the almost universal and axiomatic use of cutting scores. I'm not referring to the kind of cutting score marking the lower limit of ranked scores on an eligibility list. I'm referring to the kind of cut score above which anyone who comes can be hired and below which no one will be -- the kind that changes a continuous score distribution to a dichotomy. A major part of my frustration is with the reason most often given for setting cut

scores: "My managers just can't handle anything more complicated than a pass-or-fail score."

Fervent wish #11: I wish I knew why and when we stopped assuming that decision makers had any brains. If they are as unteachable as many psychologists assume, we certainly won't get very far with some of the other items on my wish list. For example, we couldn't hope to teach them to use a specified policy for combining inconsistent predictions and would have to stay with the same tired old overall criterion.

12. Most of all, I wish I knew how to handle the people who don't seem to realize that assessments will be made, that they will be better if competently made, and that decisions have to be made about the way in which opportunities are to be allocated in society. You will surely recognize the frustrations behind this item in my list. We face daily a society and its government that, on the one hand, insists on testing for licensure, for diplomas, and sometimes for scholarships and, on the other hand, insists that tests are inherently invalid, biased, and violations of civil rights. We face a bill in Congress that is fundamentally directed to tests as instruments of employment decision making but has no corresponding concern about interviews. We get nonsense from adversarial lawyers who seek to solve no problem but to win their kinds of cases. Consider the recent Harvard Law Review paper by Kelman (1991). He argued that what he called general ability tests are necessarily invalid, even statistically, because they do not correlate well with pay, "the best presumptive measure of worker productivity in a market system" (Kelman, 1991, p. 1208). Even for a "perfectly valid" test that is "racially unbiased in the Cleary sense," "It is still not obvious that those with higher test scores are entitled

to the jobs that by hypothesis they are better able to perform" (Kelman, 1991, p. 1243, italics in original). We get reports like the recent one from the "National Commission on Testing and Public Policy" that complains of the overuse of multiple choice examinations without a systematic evaluations of alternatives to them. (Kelman, at least, offered an alternative to testing: hire those who apply to see how well they can eventually perform -- with, of course, curbs on the right of employers to fire those deemed poor performers.)

Somehow, we need to get the questions asked in the right order. The first question is not whether to test, or how to test, but how society's opportunities and rewards are to be allocated. We have generally said that nepotism is not an acceptable basis, we disparage employers who choose on the basis of applicant appearance, and we pass laws against the allocation of rewards on the basis of sex, race, or ethnic identification. We oppose quotas; we promote affirmative action; unfortunately, we don't articulate the difference very well. We aren't totally sure, but we think we don't like random allocation, not even of the first come, first hired variety. We like to say, at least some of us do, that we think opportunity should be allocated on the basis of merit. Only if we agree on that do we reach the issues of testing. And then we must ask how merit is best identified. One way to assess merit is to test. Another way is to interview. Still another is Kelman's open-ended probationary period.

We can stipulate that, if merit is to be assessed by testing, the testing must be done competently, and we can probably agree on at least some of the requirements of competence. But there are still questions. There are many problems with testing, and those of us who are here

probably know those problems better than the critics do. But when we've considered the problems, we still have one remaining question: "What will do the job better?" How can we get this sequence of questions to characterize the deliberations of those concerned with public policy? How can we get such deliberations to be directed toward solving a specifiable problem rather than to trumpet previously held adversial positions? I wish I knew.

References

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Chicago: Rand-McNally.

Dunnette, M. D. (1963a). A modified model for test validation and selection research. Journal of Applied Psychology, 47, 317-323.

Dunnette, M. D. (1963b). A note on the criterion. Journal of Applied Psychology, 47, 251-254.

Guion, R. M. (1961). Criterion measurement and personnel judgments. Personnel Psychology, 14, 141-149.

Guion, R. M., & Gottier, R. F. (1965). Validity of personality measures in personnel selection. Personnel Psychology, 18, 135-164.

Kelman, M. (1991). Concepts of discrimination in "general ability: job testing. Harvard Law Review, 104, 1158-1247.

National Commission on Testing and Public Policy. (1990). From Gatekeeper to gateway: Transforming testing in America. Chestnut Hill, MA: Author, Boston College.

Van Zelst, R. H. (1952). Validation of a sociometric regrouping procedure. Journal of Abnormal and Social Psychology, 47, 299-301.

Figure 1

Satisfies Policies?
No
Reject
Yes
Interview
Prior Accomplishments?
Job Experience
(Yes or No)
None; Yes
Reject
None; No
Some; No
Assess Aptitude, Motivation
Good; hire
Poor; reject
Many; No
Many; Yes
Some; Yes
Hire or Consider for Other Jobs
Work Sample
High Score
Low Score